

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-067355  
 (43)Date of publication of application : 16.03.2001

(51)Int.Cl. G06F 17/27  
 G06F 17/21  
 G06F 17/28  
 G06F 17/30

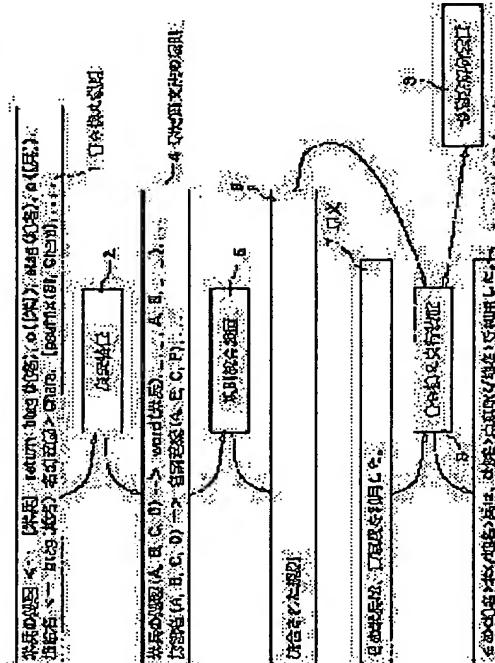
(21)Application number : 11-241245 (71)Applicant : NIPPON TELEGR & TELEPH CORP  
 <NTT>  
 (22)Date of filing : 27.08.1999 (72)Inventor : ISOZAKI HIDEKI

(54) SENTENCE REWRITING/INFORMATION EXTRACTING METHOD WHICH ENABLES CONSTRAINED DESCRIPTION OF CHARACTER STRING, DEVICE THEREFOR AND RECORDING MEDIUM

## (57)Abstract:

**PROBLEM TO BE SOLVED:** To briefly and declaratively specify a rewriting rule or an inherent expression extraction rule including a constraint of a character string without caring about order of processing in a system which summarizes a huge amount of document information, corrects an expression into an audible one, on the contrary, transforms a sentence written in spoken word into a written word that is easier to read, and extracts important structural elements (inherent expression) such as a person's name, a place name, a name of an organization, or time and date from a newspaper article or the like.

**SOLUTION:** This document rewriting transforms a set 1 of rewriting rules a user describes into a set 4 of rules of established phrase grammar by a translating device 2 and then transforms the set 4 of the rules of established phrase grammar into an integrated rule 6 which can perform high speed parallel processing. A rewriting execution device 8 receives the integrated rule 6 and a document (an original sentence 7) to be transformed and outputs a transformed result 9.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(書誌十要約十精求の範圍)

妻分なるよつに表現に於ける成する各種の特徴等である。

【001】ここで、ボディー部に return による書き換え指定がある場合には、その return より前の部分で得られる出力を無視し、return の後に指定された文字列を連結して得られる文字列の差分リスト表現が、ヘッド部の出力用の引数の差分リスト表現になるように、差分リストを構成する各変数を設定する。

[0001]

[0012]なお、ボディー部に現れる終端記号は、対応する文字列を出力側にコピーするための述語の呼出しに置き換える。また、ボディー部に現れる付加制約は、そのまま翻訳後の規則に残す。  
[0013]以上が、翻訳ステップの具体的な内容である。  
[0014]本発明の文章書き換え・情報抽出装置は、ユーザが与えた書き換え規則や抽出規則を論理型言語で実行可能な形式に変換するための翻訳手段と、複数の規則を統合して高速処理の可能な形式に変換するための統合手段と、さらにに統合された規則を実際の文章に適用するための適用手段と有する。

[0015] (作用) 本発明によれば、ユーザが書き換える規則あるいは抽出規則の集合をシステムに与えると、システムはこれを論理型言語で実行可能な形式に変換し、これら規則を統合してさらに高速処理の可能な形式に変換し、統合後の規則を用いて実際に二章の書き換え処理あるいは情報抽出処理を行なうので、ユーザは、文字列レベルの判断を含む文書をもとにした処理を簡単に実行することができる。

[0016]

【発明の実施の形態】次に、本発明の好ましい実施の態様について、図面を参照して説明する。

2 図示す。図の二種類の発明本発明

は、図1に示す装置を使用し、本発明に基づく文章書き換え・情報抽出方法を実施する際の処理手順を示すデータ

[0018] 図1に示す装置は、ユーザが与えた書き換え規則や抽出規則を論理型言語で実行可能な形式に変換する機能である。

ある規則封合装置55と、結合された規則を複数の文書に適用するための通用手段で構成される。規則封合装置55は、各規則を複数の文書に適用するための統合手段である。

換え実行装置8により文章の書き換えを行う際に文書の形態素解析を行う形態素解析装置3と、各規則を格納する

規則別冊部11と、書き換え装置8に与える文章を格納する出力文書格納部13とを備えている。規則格納部11には、具体的には、ユーザが指定した書き換え規則または地

2つで結合された後の規則(6図2)を終結する。  
出規則)の集合1(図2)と、翻訳装置2で変換された後の確定節文法の規則の集合4(図2)と、規則統合装置5によ

[0019]次に、この文章書き換え・情報抽出接続を用いた文章書き換え・情報抽出について、図2のデータフロー図

[0020]まず、船艤装置2が、ユーザが指定し規則格納部11に格納されている書き換え規則(あるいは抽出規則)集を用いて読みます。

規則合4を確定筋文法での規則の集合4に変換する。さらに確定筋文法では規則4は、規則統合装置5によって高層物理理の可能な形式で統合された規則6に変換される。

納部12内の原文7を書き換えるのに利用され、書き換えた結果9は、出力されて出力文書格納部13内に格納され

[0021]なお、ユーザの与える規則は、論理型言語で文法解析を行なう場合に一般的に用いられている「確定筋文」として扱われる。

法 (DCG: Definite Clause Grammer) に準じた記法を用いるものとする。ただし、通常の確定節文法が入力に関する記述を記述する記述から外して、文書中の記述を記述する記述から外す。  
出力は、  
1. 確定節文法の記述  
2. 文書中の記述

[0022]ます、名詞がひとつ以上連続して出現した場合にマッチする「名詞連接」という概念は、確定節文法を用いた

場合と同様に、以下のように定義できる。

【数1】名詞連続 <.. 名詞、(名詞連続)1.これは、「名詞」の直後に「名詞連続」がまたはになければ、それ全体を

名前が運営を貢献させることを表す「貢献」を用いて、規則は「貢献規則」と名づけられています。この規則は、出力がそのまま出力にコピーされるものとする。したがって、ここで用いる規則では、出力について記述のない規則は、入力がそのまま出力にコピーされるものとする。したがって、ここ

[0024] 次に、固有表現抽出の例について二つの名詞連接に対する入力を文書別ヒートマップ図会に示す。この実験においては、そのまま出力へコピーされる。

「駅」や「会館」などの施設を示す文字列であって、「特急通過駅」などの普通名詞的なものでなければ、全体を「地名」

【数2】 施設名 <...> 全国連携システム(SfChango) 基盤のサブシステム(Su) 普通名目的な施設

(Chars), etag(地名). まず、最初の「etag(地名)」と最後の「etag(地名)」は、入力がこの規則とマッチした場合に、出

力のこの位置に、それぞれ「<地名>」「<地名>」というタグが挿入されることを示す(htag(x)は、xについての開始タグ「<x>」を挿入することを表し、etag(x)は、xについての終了タグ「</x>」を挿入することを表している)。「名詞連続Chars」は、上記で定義した「名詞連続」とマッチした部分の入力の文字列を、この規則では CharS と呼ぶことに意味している。次の「[...]」の部分は、この文字列 CharS に関する制約を論理型言語で実行可能な形式で表してい る。具体的には、文字列 CharS のうらの方(サフィックス)に、「駅などの施設を表す文字列 ST」が付いていて、しかも CharS が「特急通過駅」のような普段名詞的なものではない、という制約を表している。つまり、この規則では、部分は制約を説述しているだけであり、出力には影響しない。

〔002〕なお、ここで用いられた制約を記述するために用いた記号類は、実装に用いる論理型言語を使って、次のように直接定義することができる。

(0027)施設のサフィックス([記]).施設のサフィックス([記]).普通名詞的な施設([急,通,駅]).これらの制約は単語列ではなく、文字列に関するため、書き換える実行装置8が形態素解析装置3を呼出して原文7の形態素解析を行わせた際の、原文7をどのように文章を分割するかに依存しない、という特徴を有する。ここでは例えば、これらの制約をたとえれば、次のようになります。上記の規則もこれに合わせて書き換えたとしよう。

[0030] [数3]米兵の処理 <.. [米兵] return btag(地名,o([米]),etag(地名),o([兵]).最初の「[米兵]」は、「米兵」という単語が入力に現れた場合という制約を表わしている。規則中に return があると、システムはその左側に書かれた入力を取出しにコピーすることはせず、return のあとに書かれた指示に従い出力する。ここでは o(..)によって、「米」という文字「[兵]」という文字を別々に出力することを指示している。これによって、「<地名>米</地名>兵」と書き換えたならば、以下のように書けばよく、本発明の方法が固有表現抽出だけでなく、一般の書き換えにも使えることが分かる。

[0031]もし、「米兵」を「<地名>アメリカ</地名>兵」と書き換えたならば、以下のように書けばよく、本発明の方法

【数4】  
[米兵]の処理 ← [米兵]return btag(地名).o([アメリカ]), etag(地名).o([兵])翻訳装置2は、以上のような形式によりユーザーによって書かれた誤用を読み込んで、論理型言語で実行可能な確定文法の規則に変換する。上記の各規則は以下のようないくつかの規則に変換される。

[0035]「word(W,F,I,P2,A,B,C,D)」は、Wという単語がその場所に出現したことを表わす。「F1」や「F2」は形態素解析装置3が出力した単語Wの特徴であり、この例では、F1が漢字やカタカナといった文字の種類、F2が品詞や活用形などの情報を記録している。したがって、本実施形態での規則に「word(…,カタカナ)」と書いておけば、任意のカタカナ単語とマッチする。また、「word(…,接尾辞)」と書いておけば、任意の接尾辞とマッチする。逆に「[米兵]」は「word(米兵,+)」の略語と見なせる。確定文法文の規則に変換されたときに加わる word の最後の4つの引数は、非終器記号の場合と同じく、入力の文字列を出力へ自動的にコピーする作業を行なう機能を有する。つまり、「word(T,F1,F2,A,B,C,D)」は、単語 T の文字列表現を入力側から出力側にコピーするため、 $diff(A,B,L),append(L,D,C)$ を実行する。たとえば、Tが「東京駅」であれば、入力の文字列表現を示す A は「[東京駅]」である。word の  $diff(A,B,L),append(L,D,C)$  の処理によって、Cが「[東京駅,D]」になる。この自動的なコピー機能がはじめから組み込まれているため、本実施の形態の方法では、出力に関する記述が簡単である。

[0036]なお、word は、実際に、入力文字列を出力へ自動的にコピーするため、 $word(T,F1,F2,A,B,C,D)$  は、 $word(T,F1,F2,A,B,C,D)$  と書く。形態素解析装置3が出手した単語と品詞情報のリストを要けける引数は、これらの確定文法文上には明示されていないが、通常の確定文法文に従い、引数の最後に差分リストの形で加わる。つまり、たとえば「施設名(A,B,C,D,E,F)」のように、形態素解析の結果を受け取るために引数 E,F が加えられる。

[0038]以上のようにして得られた確定筋文法の規則4は、上昇型チャート法などの既存手法を用いることにより、すべての可能性を効率よく並行して計算することができる。そこで規則組合装置5が、確定筋文法の規則4を効率よく並行計算できる規則群に変換し、書き換え実行装置8にこれを実際に入力する。[0039]なお、複数の書き換え方がある場合は、書き換え実行装置8がその中から最適なものを一つ選び出して適用する。最適なもののが判断基準では、たとえば入力が先頭から何回あるかは単純な比較がマッチするがマッチするかを比べて、どちらでも最も長くマッチするがマッチするかが選ぶ、いわゆる最長一致法を用いることができる。

【0041】  
【数6】  
Head <- InputSeq return outputSeq Head <- InputSeq Head .. InputSeq ここで、Head は一つの原子式である。また InputSeq には、1)終端記号、2)非終端記号、3)終端記号または非終端記号の列に文字リスト化演算子の付いたもの、4)実戦に用いた論理型値の直書き呼び出し、5)出力コマンド、6)入力コマンド、7)出力リストである。文字リスト化演算子はたとえば G>L のよう [0042] 上記の実戦例では、A または B であることを示す(A,B)という表記が用いられているが、処理の説明が必要以上に複雑になるため、ここでは考えない。このような規則は A の場合と B の場合の2つの規則に分けて説明するこ

[00441] 替き換え規則をひとつ説む。

【0046】3. 入力を処理する部分をまとめる場所 INPUTPART と、出力を処理する部分をまとめる場所 OUTPUTPART の間に渡される。

OUTPUTPARTを用意し、それぞれ空にしておく。  
[0047] 14：規則が return を含まない規則の場合は、後述する手続 `trans(A,C,InputSeq)` を呼び、最終的に得られる

return を含む規則の場合は、以下のように入力側の処理と出力側の処理を分けて行なう。

[0049]5-1. まず、入力側の処理のため、手続き `trans(A..Inputseq)`を呼び、手続き `trans` の処理の最後の段階で、`B` と一致させるところで、`(i) わかる箇名本数である、つまり、新しい論理節であるが、その値を参照する`

必要がないので、名前を付けていない。これは、InputSeq を出力に使わないからである。  
[0050]5-2. 次に、出力側の処理のため、outpuiseq に対して手続き transC.C.OutputSeq()を呼び、手続き  
trans の処理の最後の VO の値を D と单一化する。OutputSeq は入力側に匿名変数  
を利用している。

する規則の本体とする。そして手書き  $\text{trans}(A\_InputSeq)$  の処理の最後の VI の値を B、手続き  $\text{trans}(C\_OutputSeq)$  の最後の VO の値を D と同一化する。

[0052] 手書き  $\text{trans}(A, C, Seq)$  次に、手書き  $\text{trans}(A, C, Seq)$  について説明する。手書き  $\text{trans}(A, C, Seq)$  では、引数の受渡しを行うため、入力の現在の先頭を委譲するためのローカル変数 VI と出力の現在の先頭を表す論理変数名を記憶するためのローカル変数 VO を準備する。VI の初期値を A、VO の初期値を C とする。あとは、以下の処理 1～5 を Seq が空になるまで繰り返す。

[0053] 1. seq の先頭の要素 E が終端記号 [] であるとき、T をそのまま出力側にコピーするため、新しい論理変数 NI と NO を用意して、word(T, ..., VI, NI, VO, NO) という記号を INPUTPART の末尾に連結する。NI, NO は、T と次の単語の境界の場所を指す論理変数として使われる。そして、VI, VO の新しい値として NI, NO を採用し、E を InputSeq から削除する。

[0054] 2. Seq の先頭の要素 E が非終端記号 NT であるとき、新しい論理変数 NI と NO を用意して、NT に引数 VI, NI, VO, NI を加えたものを INPUTPART の末尾に連結する。NI, NO は、NT と次の単語の境界の場所を指す論理変数として使われる。そして、VI, VO の新しい値として NI, NO を採用し、E を InputSeq から削除する。

[0055] 3. Seq の先頭の要素 E が文字リスト化演算子の付いたもの G つとしてあるとき、まず手続き  $\text{trans}(VI, YO, G)$  を呼ぶ。さらに手書き  $\text{trans}$  を呼び出す時点で VI の値を VIa、 $\text{trans}$  が終了した時点で VI の値を VIb として、(diff(VIa, VIb), L) を INPUTPART の末尾に連結する。

[0056] 4. Seq の先頭の要素 E が実装に用いた論理型直角座標の直接呼び出し(G)である場合、G をそのまま INPUTPART の末尾に連結する。これはこれまでの入力に関する制約の追加にすぎないので、VI, YO の値は変化させない。

[0057] 5. Seq の先頭の要素 E が出力コマンド o(X) である場合、新しい論理変数 NO を用意して、OUTPUTPART の末尾に( $\text{append}(x, NO, YO)$ )を追加する。同様に、beag(X)の場合も(YO = [<X> | NO])を、etag(X)の場合も(YO = [<X> | NO])を OUTPUTPART の末尾に追加する。そして YO の新しい値として NO を採用する。これらは出力だけに関するもので、入力には関係ないので、VI の値は変化させない。

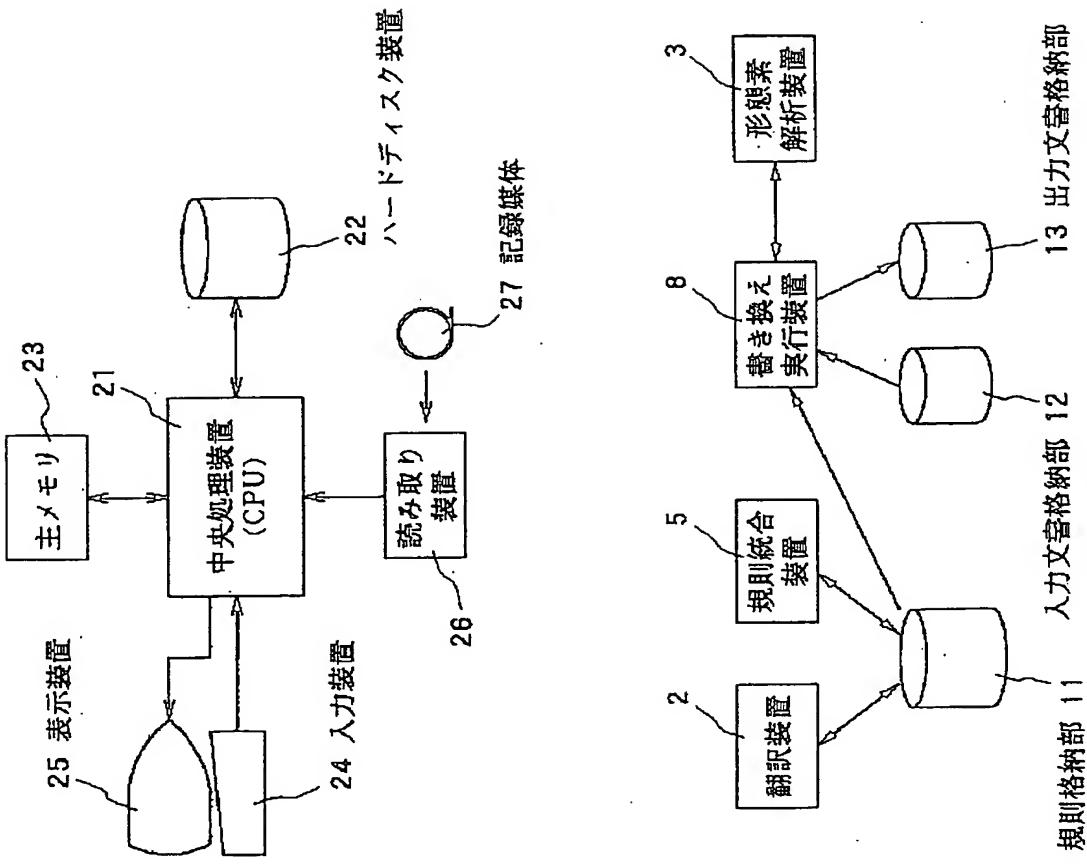
[0058] 6. 以上により、手続き  $\text{trans}(A, C, Seq)$  の処理が終了する。

[0059] 以上、本発明の特徴的な実施の形態について説明したが、上述した文章書き換え・情報抽出装置は、それを実現するための計算機プログラムを、スーパーコンピュータやエンジニアリングワークステーション(EWS)などの計算機に読み込みませ、そのプログラムを実行させることによっても実現できる。文章書き換え・情報抽出を行うためのプログラムは、磁気テープやCD-ROMなどの記録媒体によって、計算機に読み込まれる。図3は、上述の文章書き換え・情報抽出方法を実行する計算機の構成を示すブロック図である。

[0060] 図3に示した計算機は、中央処理装置(CPU)21と、プログラムやデータを格納するためのハードディスク装置22と、主メモリ23と、キーボードやマウスなどの入力装置24と、CRTなどの表示装置25と、磁気テープやCD-ROM等の記録媒体27を読み取る読み取り装置26とから構成されている。ハードディスク装置22、主メモリ23、入力装置24、表示装置25及び読み取り装置26は、いずれも中央処理装置21に接続している。この計算機では、文章書き換え・情報抽出を行うためのプログラムを格納した記録媒体27を読み取り装置26に接続し、記録媒体27からプログラムを読み出してハードディスク装置22に格納し、ハードディスク装置22に格納されたプログラムを中央処理装置21が実行することにより、文章書き換え・情報抽出が実行される。

[0061] 図1に示した文章書き換え実行装置8は、中央処理装置21におけるプログラムの実行により実現でき、また、規則統合装置11、入力文書格納部12および出力文書格納部13は、中央処理装置21に接続される。

[0062] [発明の効果] 以上説明したように、本発明は、膨大な文書情報を要約したり、音声で聞き取りやすい表現にしたり、逆に話し言葉で書かれた文章を読みやすい書き言葉にしたり、新聞記事等から人名や地名や組織名や日時などの文章の重要な構成要素(固有表現)を抽出したりするシステムにおいて、文字列の制約を含む書き換え規則や抽出規則を、処理の順序を気にすることなく、直観的かつ簡潔に指定することができる効果がある。



米兵の処理 <— [米兵] return btag(地名), o([米]), etag(地名), o([兵]).  
施設名 <— btag(地名), 名詞連続> Chars, [psuffix(Sf, Chars), ...]

翻訳装置

1 書き換え規則

米兵の処理 (A, B, C, D) —> word(米兵), ..., A, B, ..., ...  
施設名 (A, B, C, D) —> 名詞連続 (A, E, C, F), ...

4 確定節文法の規則

規則統合装置

5

統合された規則

7 原文

その米兵は、東京駅を利用した。

書き換え実行装置

6

その<地名>米</地名>兵は、<地名>東京駅</地名>を利用した。

3 形態素解析装置

9 書き換えられた文

8